Bases de Données

Amélie Gheerbrant

Université Paris Diderot
UFR Informatique
Laboratoire d'Informatique Algorithmique : Fondements et Applications

amelie@liafa.univ-paris-diderot.fr

4 novembre 2014

Le Modèle Relationnel

- Les données sont organisées dans des relations (tables)
- Schéma de bases de données relationnel
  - ensemble de noms de tables
  - liste d'attributs pour chaque table
- Les tables sont spécifiées sous la forme :

  `<nom de la table> : <liste d'attributs>`
- Exemples :

  `Compte: numero, agence, clientId`

  `Film: titre, directeur, acteur`
- Dans une même table les attributs ont des noms différents
- Les tables ont des noms différents

Exemple de base de données relationnelle

Film	Titre	Réalisateur	Acteur
	Shining	Kubrick	Nicholson
	The Player	Altman	Robbins
	Chinatown	Polanski	Nicholson
	Chinatown	Polanski	Polanski
	Repulsion	Polanski	Deneuve

Projection	Cinéma	Titre
	Le Champo	Shining
	Le Champo	Chinatown
	Le Champo	The Player
	Odéon	Chinatown

Exemples de requêtes

- Trouver le nom des films projetés en ce moment :

réponse	titre
	Shining
	The Player
	Chinatown

- Trouver les cinémas qui passent des films de Polanski :

réponse	cinéma
	Le Champo
	Odéon

- Trouver les cinémas qui passent des films avec Nicholson :

réponse	cinéma
	Le Champo
	Odéon

- Trouver tous les réalisateurs qui se sont dirigés eux-mêmes :

réponse	director
	Polanski

- Trouver tous les réalisateurs dont les films sont joués dans tous les cinémas :

réponse	cinéma
	Polanski

- Trouver tous les cinémas qui ne passent que des films avec Nicholson :

réponse	cinéma

mais si le Champo cesse de passer 'The Player', la réponse devient :

réponse	cinéma
	Le Champo

Les Résultats des Requêtes

- Ce sont des tables construites à partir d'autres tables de la base de données

Comment formuler une requête ?

- Deux types de langages de requêtes :
  - Commercial : SQL
  - Théorique : le Calcul Relationnel, l'Algèbre Relationnelle, Datalog, etc

Déclaratif versus Procédural

Déclaratif :

$f$cinéma j (titre, réalisateur, acteur) *2* film,
(cinéma, titre) *2* Projection,
acteur='Nicholson' g

Procédural :

```
for each tuple T1=(t1,r,a) in relation film do
    for each tuple T2=(c,t2) in relation projection do
        if t1=t2 and a='Nicholson' then output c
    end
end
```

Déclaratif versus Procédural

- **Langages théoriques :**
  - Déclaratif : le Calcul Relationnel, les requêtes basées sur des règles
  - Procédural : l'Algèbre Relationnelle

- **Langages utilisés en pratique :** mélange des deux, mais surtout déclaratif

On va voir quoi aujourd'hui ?

- exemple de requêtes dans di□érents langages

Exemples de requêtes

- Trouver le nom des films projetés en ce moment :

  `réponse(ti) :- film(ti, real, act)`

  i.e.,

  while tuple `(ti, real, act)` dans la relation film,
  output `ti` (valeur de l'attribut titre)

- Formulation des requêtes comme règles indiquant quand certains éléments appartiennent à la réponse.

- On appelle ça des requêtes conjonctives (pourquoi ? +tard)

Autre exemple

- Trouver les cinémas qui passent des films de Polanski :

`réponse(ci) :- film(ci, 'Polanski', act), projection(ci, ti)`

i.e.,

while `(ti, real, act)` dans la relation film,
tester : `real='Polanski'?`;
si non, considérer le tuple suivant,
si oui, considérer tous les tuples `(ci,ti)` dans Projection,
et pour chacun, *output* `ci` (valeur de l'attribut cinema)

= type de requête le plus répandu.

- Trouver les réalisateurs qui se sont dirigés eux-mêmes :

```
réponse(real) :- film(ti, real, act), real=act
```

i.e.,

while `(ti, real, act)` dans la relation film,
tester : `real=act?` ;
si non, considérer le tuple suivant,
si oui, `output real`.

Un exemple plus compliqué

Trouver les réalisateurs dont les films passent dans tous les cinémas...

- "Tous" : souvent problématique.
- Besoin du quantificateur universel *8*

*f*real *j* *8* (ci, ti) *2* Projection, *9* (ti', act) : (ti', real, act) *2* Film ^ (ci, ti') *2* Projection *g*

pour tester si `real` *2* réponse, pour tout nom de cinéma `ci`, tester s'il existe un tuple `(ti', real, act)` dans Film et un tuple (ci, ti') dans Projection.

Notation de la logique mathématique :

- *8* signifie "pour tout", *9* signifie "il existe"
- ^ est une conjonction (ET logique)

SQL, les raisons du succès

- SQL = *Structured Query Language* (IBM fin 1970')
- Standards : SQL-86, SQL-92, SQL-99 / SQLL3 (+1000 pages)
- Requêtes basées sur le modèle relationnel : langage logique, simple et compréhensible.
- Une requête du calcul peut être facilement traduite en une expression de l'algèbre qui s'évalue simplement (Théorème de Codd)
- Algèbre relationnelle = modèle limité de calcul (n'autorise pas les fonctions arbitraires). Autorise l'optimisation de l'évaluation des expressions algébriques.
- Parallélisme facilité pour les très grandes bases de données.

Exemples de requêtes SQL

- Trouver le nom des films projetés en ce moment :

```
SELECT Titre
FROM FILM ;
```

- SELECT liste les attributs retournés par la requête
- FROM liste les relations prises en entrée

Plus d'exemples

- Trouver les cinémas qui passent des films de Polanski :

```
SELECT Projection.cinema
FROM Projection , Film
WHERE Film.titre = Projection.titre
                AND Film.realisateur = 'Polanski' ;
```

Di□érences :

- SELECT spécifie maintenant de quelle relation viennent les attributs - parce qu'on en utilise plus d'une
- FROM liste deux relations
- WHERE spécifie les conditions de sélection des tuples

Jointures de relations

- WHERE autorise à faire la jointure de plusieurs relations.

Requête : lister les réalisateurs avec les cinémas dans lesquels passent leurs films

- Requête conjonctive :

```
réponse(real, ci) :- Projection(ci, ti), Film(ti, real, act)
```

- Requête SQL :

```
SELECT Film.realisateur, Projection.cinema
FROM Projection, Film
WHERE Film.titre = Projection.titre;
```

Jointures de relations

- ```
  SELECT Cinema.realisateur, Projection.cinema
  FROM Projection, Film
  WHERE Film.titre = Projection.titre;
  ```

- Sémantique : boucles imbriquées sur les relations listées dans le `FROM`

```
for each tuple (titre1, réalisateur, acteur) in Film do
     for each tuple (cinema, titre2) in Projection do
           if titre1=titre2 then output (realisateur, cinema)
     end
end
```

- Cette opération s'appelle une jointure : une des opérations les plus fondamentales en BD.
```

Un langage procédural : l'algèbre relationnelle

- Commençons par un sous-ensemble de l'algèbre relationnelle qui su□t à capturer les requêtes simples basées sur les règles et les énoncés SQL simples de la forme `SELECT-FROM-WHERE`.

- Ce sous-ensemble a trois opérations :

  Projection π
  Sélection σ
  Produit Cartésien

- Parfois on utilise aussi le renommage ρ, mais on peut l'éviter sous certaines conditions.

La Projection

- Choisit des attributs dans une relation.
- $\pi_{A_1,\ldots,A_n}(R)$: conserve uniquement les attributs $A_1, \ldots, A_n$ dans la relation R
- Exemple

$\pi_{titre,realisateur}$

titre	réalisateur	acteur
Shining	Kubrick	Nicholson
The Player	Altman	Robins
Chinatown	Polanski	Nicholson
Chinatown	Polanski	Polanski
Repulsion	Polanski	Deneuve

=

titre	réalisateur
Shining	Kubrick
The Player	Altman
Chinatown	Polanski
Repulsion	Polanski

- Fournit à l'utilisateur une vue des données en omettant certains attributs

La sélection

- Choisit des tuples satisfaisant certaines conditions.
- $\sigma_{cond}(R)$: conserve uniquement les tuples t pour lesquels la condition $cond(t)$ est vraie.
- Conditions : conjonctions de
  - `R.A = R.A'` deux attributs ont la même valeur
  - `R.A = c` la valeur de l'attribut est la constante c
  - Idem mais avec $\neq$ à la place de $=$
- Exemples :
  - `Film.acteur = Film.realisateur`
  - `Film.acteur` $\neq$ `Nicolson`
  - `Film.acteur = Film.realisateur ^ Film.acteur = Nicolson`
- Fournit à l'utilisateur une vue des données en omettant les tuples qui ne satisfont pas certaines conditions voulues par l'utilisateur.

La sélection : exemple

$\sigma_{acteur=realisateur \wedge realisateur='Polanski'}$

titre	réalisateur	acteur
Shining	Kubrick	Nicholson
The Player	Altman	Robins
Chinatown	Polanski	Nicholson
Chinatown	Polanski	Polanski
Répulsion	Polanski	Deneuve

=

titre	réalisateur	acteur
Chinatown	Polanski	Polanski

Combiner sélection et projection

- Trouver les réalisateurs qui ont joué dans leurs propres films :
- `réponse(real) :- film(ti,real,act), act=real`
- ```
  SELECT realisateur
  FROM Film
  WHERE realisateur=acteur ;
  ```
- Requête de l'algèbre relationnelle :

$$Q = \pi_{realisateur}(\sigma_{realisateur=acteur}(\text{Film}))$$

- $\sigma_{realisateur=acteur}(\text{Film})$ donne

titre	réalisateur	acteur
Chinatown	Polanski	Polanski

- D'où $\pi_{realisateur}(\sigma_{realisateur=acteur}(\text{Film}))$ donne

réalisateur
Polanski
  ```

Combiner sélection et projection

- Il peut y avoir plusieurs façons d'écrire la même chose
- Exemple : trouver les films et les directeurs en excluant les films de Polanski
- `réponse(ti,real) :- film(ti,real,act), real` $\neq$ `'Polanski'`
- Requête de l'algèbre relationnelle :

$$Q_1 = \sigma_{realisateur \neq `Polanski'}(\pi_{titre,realisateur}(Film))$$

- Une autre requête, équivalente, de l'algèbre relationnelle :

$$Q_1 = \pi_{titre,realisateur}(\sigma_{realisateur \neq `Polanski'}((Film))$$

- La même requête déclarative peut avoir plusieurs traductions procédurales.

Combiner sélection et projection

- Q_1 et Q_2, est-ce que c'est *la même* requête ?
- Sémantiquement, oui : elles produisent le même résultat
- Mais elles di□érent en termes d'e□cacité.
- Q_1 parcourt d'abord `Film`, projette deux attributs, et parcourt à nouveau le résultat.
- Q_2 parcourt `Film`, sélectionne certains tuples, et parcourt ensuite les tuples sélectionnés.
- Q_2 semble donc e□cace dans le contexte présent.
- Les langages procéduraux peuvent être optimisés : il y a des manières sémantiquement équivalentes d'écrire la même requête, et certaines sont plus e□caces que d'autres.

Le Produit Cartésien

- Si R_1 a n tuples et R_2 a m tuples, alors $R_1 \quad R_2$ a $n \quad m$ tuples
- Opération coûteuse : si R et S ont chacun 1000 tuples (petites relations), $R \quad S$ a 1 000 000 tuples ($6=$ petit).
- Les algorithmes d'optimisation de requêtes essaient d'éviter la construction des produits - à la place ils tentent d'en construire seulement des sous-ensembles ne contenant que les informations pertinentes.

Le Produit Cartésien : exemple

Trouver les cinémas qui jouent des films de Polanski :

```
réponse(ci) :- Film(ti,real,act), Projection(ci,ti), real='Polanski'
```

- Étape 1 : Soit $R_1 = Film \quad Projection$
- On ne veut que les tuples dans lesquels les titres sont identiques, d'où :
- Étape 2 : Soit $R_2 = \sigma_{Film.titre = Projection.titre}(R_1)$
- Étape 3 : On ne veut que les films de Polanski, d'où :

$$R_3 = \sigma_{realisateur = `Polanski'}(R_2)$$

- Étape 4 : Dans la réponse, on ne veut que des cinémas, donc :

$$\text{Réponse} = \pi_{cinema}(R_3)$$

- En résumé, la réponse est :

$$\pi_{cinema}(\sigma_{realisateur = `Polanski'}(\sigma_{Film.titre = Projection.titre}(Film \quad Projection)))$$

Le Produit Cartésien : exemple

- La réponse est :

$$\pi_{cinema}(\sigma_{realisateur=`Polanski'}(\sigma_{Film.titre=Projection.titre}(Film \quad Projection)))$$

- Mais plusieurs sélections peuvent être combinées en une seule :

$$\sigma_{cond_1}(\sigma_{cond_2}(R)) = \sigma_{cond_1 \wedge cond_2}(R)$$

(avec $cond_1$, $cond_2$ des conditions Booléennes sur les tuples)

- Au final la réponse à la requête est donc :

$$\pi_{cinema}(\sigma_{realisateur=`Polanski' \wedge Film.titre=Projection.titre}(Film \quad Projection))$$

SQL et l'Algèbre Relationnelle

Il nous faut maintenant traduire d'un langage déclaratif à un langage procédural.

- Idée :
  `SELECT` correspond à la projection π
  `FROM` au produit Cartésien
  `WHERE` à la sélection σ

- Cas simple : juste une relation dans le `FROM`

```
SELECT A,B,
FROM R
WHERE condition c ;
```

sera traduit

$$\pi_{A,B,\ldots}(\sigma_c(R))$$

Traduction des requêtes déclaratives dans l'algèbre relationnelle

- Trouver le titre de tous les films

  ```
  réponse(ti) :- Film(ti,real,act)
  ```

- ```
  SELECT Titre
  FROM Film ;
  ```

- C'est juste une projection :

$$\pi_{titre}(Film)$$
  ```

Exemples de Traductions

- Trouver tous les cinémas qui jouent des films de Polanski :

```
SELECT Projection.cinema
FROM Projection, Film
WHERE Film.titre = Projection.titre
AND Film.realisateur='Polanski' ;
```

- D'abord, traduire sous forme de règle :

```
réponse(ci) :- Projection(ci,ti), Film(ti,'Polanski',act)
```

- Ensuite, convertir en une règle dans laquelle :
  - les constantes n'apparaissent que dans les conditions
  - toutes les variables sont distinctes
- ce qui nous donne :

```
réponse(ci) :- Projection(ci,ti), Film(ti',real,act), real = 'Polanski', ti=ti'
```

Exemples de Traductions

```
réponse(ci) :- Projection(ci,ti), Film(ti',real,act), real = 'Polanski', ti=ti'
```

- Deux relations) produit Cartésien
- Conditions) sélection
- Sous-ensemble des attributs dans la réponse) projection

Étapes :

1. $R1 = Projection \quad Film$
2. On ne veut parler que d'un seul et même film :

$$R_2 = \sigma_{Projection.titre = Film.titre}(R_1)$$

3. On ne veut que les films de Polanski :

$$R_3 = \sigma_{Film.realisateur} = Polanski(R_2)$$

4. Dans la réponse on ne veut que les cinémas :
   réponse = $\pi_{Projection.cinema}(R_3)$

Traduction formelle : de SQL aux règles

```
SELECT liste d'attributs <Ri.Aj>
FROM R1, ..., Rn
WHERE condition c ;
```

se traduit :

```
                                  R1(<attributs>)
réponse(<Ri.Aj>)   :-             ...,
                                  R1(<attributs>)
                                  c
```

Des règles à l'algèbre relationnelle

- Comment les règles sont-elles traduites dans l'algèbre ?

  `réponse(a1,...,ak) :- R1(`$\bar{A}_1$`), ...,Rn(`$\bar{A}_n$`),conditions`

- D'abord, s'assurer que les attributs sont deux à deux distincts : si on a $R_i(\ldots, A, \ldots)$ et $R_j(\ldots, A, \ldots)$ avec $i \neq j$, alors, traduire en $R_i(\ldots, A', \ldots)$ et $R_j(\ldots, A'', \ldots)$ et ajouter $A' \neq A''$ aux conditions.

- Exemple : `réponse(ti,real) :- Film(ti,real,act), Projection(ci,ti)`

  devient

  `réponse(ti,real) :- Film(ti',real,act), Projection(ci,ti'', act), ti'=ti''`

- Ces règles sont ensuite traduites :

$$\pi_{a_1,\ldots,a_k}(\sigma_{conditions}(R_1 \quad \ldots \quad R_n))$$

Enfin, de SQL à l'algèbre relationnelle

- En combinant les traductions :
  de SQL vers les règles, puis des règles vers l'algèbre
  on obtient la traduction suivante, de SQL vers l'algèbre :

```
SELECT liste d'attributs <Ri.Aj>
FROM R1, ..., Rn
WHERE condition c ;
```

devient

$$\pi_{<R_i.A_j>}(\sigma_c(R_1 \quad \ldots \quad R_n))$$

La jointure naturelle : un exemple

Titre	Réalisateur	Acteur
Shining	Kubrick	Nicholson
The Player	Altman	Robbins
Chinatown	Polanski	Nicholson
Chinatown	Polanski	Polanski
Repulsion	Polanski	Deneuve

$\bowtie$

Cinéma	Titre
Le Champo	Shining
Le Champo	Chinatown
Le Champo	The Player
Odéon	Chinatown

=

Titre	Realisateur	Acteur	Cinema
Shining	Kubrick	Nicholson	Le Champo
The Player	Altman	Robbins	Le Champo
Chinatown	Polanski	Nicholson	Le Champo
Chinatown	Polanski	Nicholson	Odéon
Chinatown	Polanski	Polanski	Le Champo
Chinatown	Polanski	Polanski	Odéon

La jointure naturelle

- La jointure n'est pas une nouvelle opération de l'algèbre relationnelle.
- Elle est définissable à partir de $\pi, \sigma,$
- Soit R une relation sur des attributs $A_1, \ldots, A_n, B_1, \ldots, B_k$
- S une relation sur des attributs $A_1, \ldots, A_n, C_1, \ldots, C_m$
- $R \bowtie S$ a pour attributs $A_1, \ldots, A_n, B_1, \ldots, B_k, C_1, \ldots, C_m$

$$R \bowtie S$$
$$=$$
$$\pi_{A_1,\ldots,A_n,B_1,\ldots,B_k,C_1,\ldots,C_m}(\sigma_{R.A_1=S.A_1 \wedge \ldots \wedge R.A_n=S.A_n}(R \quad S))$$

Propriétés de la jointure

- Commutativité : $R \bowtie S = S \bowtie R$
- Associativité : $R \bowtie (S \bowtie T) = (R \bowtie S) \bowtie T)$
- On peut donc écrire $R_1 \bowtie R_2 \bowtie \ldots \bowtie R_n$

Requêtes Select-Project-Join (SPJ)

- Il s'agit des requêtes les plus courantes.
- Règles simples, ou requêtes SELECT-FROM-WHERE simples.
- Trouver les cinémas qui passent des films de Polanski :

```
réponse(ci) :- Projection(ci,ti), Film(ti,'Polanski',act)
```

- Comme requête SPJ :

$$\pi_{Projection.cinema}(\sigma_{realisateur=`Polanski'}(Film \bowtie Projection))$$

- En quoi est-ce une simplification de la version précédente ?

$$Projection.cinema(\ Film.realisateur=`Polanski'(\ Film.titre=Projection.titre(Film\ ./\ Projection))$$

- $\sigma_{Film.titre=Projection.titre}$ est éliminé ; car impliqué par la jointure.

Propriétés des opérateurs de l'algèbre relationnelle

Taille des résultats

- Projection : $taille(\pi(R))$ $taille(R)$
  (taille = nombre de tuples)
- Parfois, $taille(\pi(R)) < taille(R)$
- Se produit lorsque deux attributs ont les mêmes valeurs

$$\pi_{A}\left(\begin{array}{cc} A & B \\ \hline a & b1 \\ a & b2 \end{array}\right) = \begin{array}{c} A \\ \hline a \end{array}$$

- Sélection : 0 $taille(\sigma(R))$ $taille(R)$
- Dépend du nombre de tuples satisfaisant les conditions.

Taille des jointures et des produits Cartésiens

- $taille(R \quad S) = taille(R) \quad taille(S)$, mais :

$$0 \quad taille(R \bowtie S) \quad taile(R) \quad taille(S)$$

- Certains tuples peuvent ne pas participer à la jointure :

R	employee	department
	Jones	D1
	Brown	D2
	Smith	D3

$\bowtie$

S	department	office
	D1	USA
	D2	UK

$=$

$R \bowtie S$	employee	department	office
	Jones	D1	USA
	Brown	D2	UK

- $(Smith, D3)$ n'est joint avec aucun tuple de S : S ne contient pas d'information sur le département D3.

Jointure vides

Les jointures peuvent être vides :

R	employee	department
	Jones	D1
	Brown	D2
	Smith	D3

$\bowtie$

S'	department	office
	D4	France
	D5	Italy

$=$

$R \bowtie S'$	employee	department	office

Traduction : de SPJ vers les règles vers SQL

- $Q = \pi A(\sigma_c(R \bowtie S))$
- Soit $B_1, \ldots, B_m$ les attributs communs de R et S
- Requête SQL équivalente :

```
SELECT A
FROM R, S
WHERE c, R.B1 = S.B1 AND . . . AND R.Bm = S.Bm ;
```

- Règle équivalente :

```
réponse(Ā)   :-  R(<attributs de R>), S(<attributs de S>),
                 R.B1 = S.B1, ..., R.Bm = S.Bm, c
```

De SPJ vers SQL : exemple

- Trouver les réalisateurs des films joués en ce moment et dans lesquels joue Ford :

$$\pi_{realisateur}(\sigma_{acteur=`Ford'}(Film \bowtie Projection))$$

- Avec SQL :

```
SELECT Film.realisateur
FROM Film, Projection
WHERE Film.titre=Projection.titre AND Film.acteur='Ford' ;
```

Vu aujourd'hui

- Requêtes SQL simples `SELECT-FROM-WHERE`
- Mêmes requêtes données sous forme de règles
- Correspondent aux requêtes définissables dans l'algèbre relationnelle avec $\pi, \sigma,$

Sauver de l'espace

- Répéter les noms de relations plusieurs fois est un peu lourd
- SQL nous laisse donc utiliser des noms de relations temporaires pour les relations
- ```
  SELECT P.cinema
  FROM Projection P, Film F
  WHERE P.titre=F.titre AND F.realisateur='Polanski'
  ```
- Utiliser une variable après le nom d'une relation indique que la relation est temporairement renommée
  ```

Les requêtes imbriquées : un exemple simple

- Jusqu'à présent dans la clause `WHERE` nous avons utilisé des comparaisons d'attributs
- En général, une clause `WHERE` peut contenir une autre requête, et tester une relation entre un attribut et le résultat d'une autre requête.
- On parle de requêtes imbriquées, car elles utilisent des sous-requêtes
- Exemple : trouver les cinémas qui passent des films de Polanski :

```
SELECT Projection.cinema
FROM Projection
WHERE Projection.titre IN
                    (SELECT Film.titre
                    FROM Film
                    WHERE Film.realisateur='Polanski');
```

Requêtes imbriquées : comparaison

```
SELECT Projection.cinema
FROM Projection
WHERE Projection.titre IN
      (SELECT Film.titre
      FROM Film
      WHERE Film.realisateur='Polanski');
```

```
SELECT P.cinema
FROM Projection P, Film F
WHERE P.titre=F.titre
AND F.realisateur='Polanski';
```

- Sémantique = même requête
- A gauche, chaque sous-requête réfère à une relation
- Avantage de l'imbrication : on peut utiliser des prédicats plus complexes que IN

Disjonction dans les requêtes

- Trouver des acteurs qui ont joué dans des films de Kubrick OU de Polanski
- ```
  SELECT acteur
  FROM Film
  WHERE realisateur='Kubrik' OR realisateur='Polanski';
  ```
- Est-ce qu'on pourrait définir ça avec une seule règle ?
- Non !
  ```

Disjonction dans les requêtes

- Solution : les disjonctions peuvent être représentées par un ensemble de règles :

```
réponse(act) :- Film(ti,real,act), real='Kubrick'
réponse(act) :- Film(ti,real,act), real='Polanski'
```

- Sémantique : calculer la réponse à chacune des règles, puis prendre leur union
- Syntaxe alternative en SQL :

```
        SELECT acteur
        FROM Film
        WHERE realisateur='Kubrick'
UNION
        SELECT acteur
        FROM Film
        WHERE realisateur='Polanski';
```

Disjonction dans les requêtes

- Comment traduire une requête avec des disjonctions dans l'algèbre relationnel ?
- `réponse(act) :- Film(ti,real,act), real='Kubrick'` est traduit par

$$Q_1 = \pi_{acteur}(\sigma_{realisateur=`Kubrick'}(Film))$$

- `réponse(act) :- Film(ti,real,act), real='Kubrick'` est traduit par

$$Q_2 = \pi_{acteur}(\sigma_{realisateur=`Polanski'}(Film))$$

- On traduit la requête entière par $Q_1 \ [\ Q_2$:

$$\pi_{acteur}(\sigma_{realisateur=`Kubrick'}(Film)) \bigcup \pi_{acteur}(\sigma_{realisateur=`Polanski'}(Film))$$

L'union dans l'algèbre relationnelle

- Une autre opération de l'algèbre relationnelle : l'union
  $R \ [\ S$ est l'union des relations R et S

- R et S doivent avoir les mêmes attributs.
  On a maintenant quatre opérations de l'algèbre relationnelle :

  $$\pi, \sigma, \quad , [$$

  (et bien sûr, $\bowtie$, qui est définissable à partir de $\pi, \sigma, \quad$)

- Ce fragment, est appelé algèbre relationnelle positive, ou requêtes SPJU (select-project-join-union)

Interaction des opérateurs de l'algèbre relationnelle

- $\pi_A(R \ [\ S) = \pi_A(R) \ [\ \pi_A(S)$
- $\sigma_{cond}(R \ [\ S) = \sigma_{cond}(R) \ [\ \sigma_{cond}(S)$
- $(R \ [\ S) \quad T = (R \quad T) \ [\ (S \quad T)$
- $T \quad (R \ [\ S) = (T \quad R) \ [\ (T \quad S)$

où $\bar{A} = A_1, \ldots, A_n$ est une séquence d'attributs et *cond* est une condition

Requêtes SPJU

Toute requête SPJU est équivalente à une union de requêtes SPJ.

- Il su□t de propager l'opérateur d'union
- Exemple :

$$\pi_A(\sigma_{cond}((R \quad (S \ [\ T)) \ [\ W)$$

$$=$$

$$\pi_A(\sigma_{cond}((R \quad S) \ [\ (R \quad T) \ [\ W))$$

$$=$$

$$\pi_A(\sigma_{cond}(R \quad S) \ [\ \sigma_{cond}(R \quad T) \ [\ \sigma_{cond} W))$$

$$=$$

$$\pi_A(\sigma_{cond}(R \quad S)) \ [\ \pi_A(\sigma_{cond}(R \quad T)) \ [\ \pi_A((\sigma_{cond} W))$$

Équivalences

Algèbre positive relationnelle
=
Union de requêtes SPJ
=
requêtes définies par un ensemble de règles
=
requêtes SQL `SELECT-FROM-WHERE-UNION`
=
unions de requêtes conjonctives
=
requêtes définies avec $\exists, \wedge, =$

- Question : est-ce que l'intersection est une requête SPJU ?
- i.e., étant donné R, S, avec les mêmes ensemble d'attributs, peut-on définir $R \setminus S$?

Plus sur l'union

- Relation R_1 : *pere*, *enfant*

R_1	pere	enfant
	George	Elizabeth
	Philip	Charles
	Charles	William

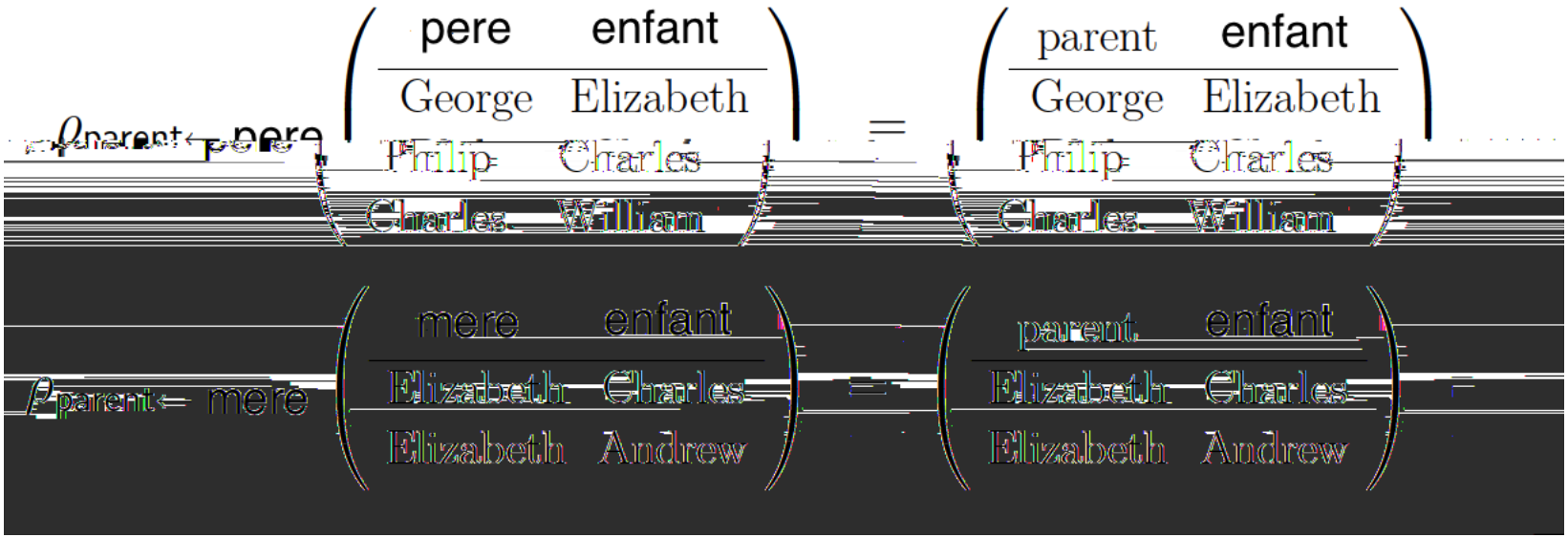
- Relation R_2 : *mere*, *enfant*

R_2	mere	enfant
	Elizabeth	Charles
	Elizabeth	Andrews

- Nous voulons leur union, qui devrait être la relation "parent-enfant"
- Mais nous ne pouvons pas utiliser R_1 [R_2, parce que R_1 et R_2 ont des attributs di□érents !
- Nous devons donc renommer les attributs

Le renommage

- Soit R une relation qui a pour attribut A, mais pas B
- $\rho_{B \leftarrow A}(R)$ est la relation qui est comme R à ceci près que A est renommé B



Le renommage

L'union désirée est :

$$\rho_{parent \leftarrow pere}(R_1) \bigcup \rho_{parent \leftarrow mere}(R_1)$$

ce qui donne

parent	enfant
George	Elizabeth
Philip	Charles
Elizabeth	Charles
Elizabeth	Andrew

SQL et le renommage

- De nouveaux attributs peuvent être introduits dans les clauses SELECT en utilisant le mot AS

```
SELECT pere AS parent, enfant
FROM R1;

SELECT mere AS parent, enfant
FROM R2;
```

- On peut prendre l'union des deux requêtes, car elles ont le même ensemble d'attributs

```
      SELECT pere AS parent, enfant
      FROM R1
UNION
      SELECT mere AS parent, enfant
      FROM R2;
```

Requêtes avec "Tous"

- Trouver les réalisateurs dont les films sont joués dans tous les cinémas

$$freal\ j\ 8(ci, ti^{0})\ 2\ Projection\ 9ti, act\ (Projection(ci, ti)\hat{}Film(ti, real, act))g$$

- Qu'est-ce que ça veut dire ?
- Pour comprendre ça il faut revenir aux requêtes basées sur les règles, et les écrire en notation logique.

Les règles revisitées

- Reprenons un exemple familier :

```
réponse(ci) :- Film(ti, 'Polanski', act), Projection(ci, ti)
```

- Qu'est-ce que ça veut dire ?
- On demande, pour chaque cinéma `ci` : "Est-ce qu'il existe un film `ti` et un acteur `act` tels que (ci, ti) appartient à `Projection` et `(ti, 'Polanski', act)` appartient à `Film` ?"
- Dans le langage de la logique mathématique, ça s'écrit comme ça :

$$Q(ci) = 9ti\,9act\ (Film(ti, 'Polanski', act) \wedge Projection(ci, ti))$$

Autres requêtes en notation logique

```
réponse(ci) :- Film(ti, real, 'Nicholson'), Projection(ci, ti)
```

est équivalente à

$$Q(ci) = 9ti\ 9real\ (Film(ti, real, 'Nicholson') \wedge Projection(ci, ti))$$

- En général, toute requête sous forme de règle unique peut être réécrite en utilisant seulement :
  - la quantification existentielle *9*
  - la conjonction logique ^ (AND)

Requêtes SPJU sous forme logique

- Trouver les acteurs qui ont joué dans des films de Polanski OU de Kubrick
- Requêtes sous forme de règles :

```
réponse(act) :- Film(ti, real, act'), real='Polanski'
réponse(act) :- Film(ti, real, act'), real='Kubrick'
```

- Notation logique :

$$Q(act) = 9ti\ 9real(Film(ti, real, act)\hat{}$$
$$(real = `Kubrick' \_\ real = `Polanski'))$$

- Nouvel élément : la disjonction logique _ (OR)
- Les requêtes SPJU s'écrivent en notation logique en utilisant :
  - la quantification existentielle *9*
  - la conjonction ^ et la disjonction _

Les requêtes avec "pour tout"

$$\{real \mid \forall (ci, ti') \in Projection, \exists ti\ \exists act (Projection(ci, ti) \wedge Film(ti, real, act))\}$$

- Nouvel élément ici : la quantification universelle $\forall$ "pour tout"
- $\forall x\ F(x) \equiv \neg \exists x\ \neg F(x)$
- Le nouvel élément est donc en fait la négation $\neg$
- Il faut faire attention avec la négation : quelle est la signification de
$$\{x \mid \neg R(x)\}$$
- Ça a l'air de nous dire : donne moi tout ce qui n'est pas dans la base de données. Mais il s'agit d'un ensemble infini !

Les requêtes avec "pour tout" et la négation

- Sûreté : une requête écrite en notation logique est sûre si elle retourne nécessairement des résultats finis sur toutes les bases de données.
- Cette propriété doit être imposée pour tous les langages pratiques.
- Mauvaise nouvelle : il n'existe pas d'algorithme pour vérifier qu'une requête quelconque est sûre.
- Bonne nouvelle : toutes les requêtes SPJ et SPJU sont sûres. Raison : tout les attributs qui occurrent dans la réponse doivent avoir une occurrence dans l'entrée, pas de création de nouvel élément.
- Problème = la négation, comment la gérer ?

Le calcul relationnel

- Variable liée : une variable x qui ocurre dans $\forall x$ ou $\exists x$
- Variable libre : une variable qui n'est pas liée.
- Les variables libres sont celles qui vont dans la réponse de la requête
- Deux façons d'écrire une requête :
  $Q(\bar{x}) = F$, où $\bar{x}$ est un tuple de variables libres
  $\{\bar{x} \mid F\}$
- Exemples :
  $\{x, y \mid \exists z\ (R(x, z) \wedge S(z, y))\}$
  $\{x \mid \exists y\ R(x, y)\}$
- Les requêtes sans variables libres sont appelées requêtes Booléennes.
- Leur réponse est toujours le *vrai* ou le *faux*. Exemples :
  $\forall x\ R(x, x)$
  $\forall x \exists y\ R(x, y)$

Le calcul relationnel sûr

- Une requête du calcul relationnel $Q(\bar{x})$ est sûre si elle retourne toujours un résultat fini.
- Exemple de requêtes sûres :
  - toute requête Booléenne
  - toute requête SPJ ou SPJU
- Exemple de requêtes non sûres :
  - $\{x \mid R(x)\}$
  - $\{x, y \mid Film(x, Polanski, Nicholson) \_ Film(Chinatown, Polanski, y)\}$
- Calcul relationnel sûr = ensemble des requêtes sûres du calcul relationnel
- Mais la sûreté ne peut pas être vérifiée par un algorithme !
- On peut tout de même décrire ce langage.

La différence

- Si R et S sont deux relations avec le même ensemble d'attributs, alors $R \quad S$ est leur di□érence : l'ensemble des tuples qui occurrent dans R mais pas dans S

- Exemple :

A	B
a1	b1
a2	b2
a3	b3

$-$

A	B
a2	b2
a3	b3
a4	b4

$=$

A	B
a1	b1

L'algèbre relationnelle

- Inclut les opérateurs $\pi, \sigma, \quad , [, \quad , \rho$

  Théorème fondamental de la théorie des bases de données relationnelles :
  Le calcul relationnel sûr = l'algèbre relationnelle

- On ne va pas donner une preuve formelle ici, mais essayer d'expliquer pourquoi c'est vrai.

Du l'algèbre relationnelle au calcul relationnel sûr

- Montrer que l'algèbre relationnelle (qui est sûre) peut être exprimée dans le calcul relationnel
- Chaque expression e produisant une relation à n attributs est traduite par une formule $F_e(x_1, \ldots, x_n)$
- R est traduit par $R(x_1, \ldots, x_n)$
- $\sigma_{cond}(R)$ est traduit par $R(x_1, \ldots, x_n) \wedge cond$

  Exemple : si R a pour attributs A et B, $\sigma_{A=B}(R)$ est traduit par $(R(x_1, x_2) \wedge x_1 = x_2)$

Du l'algèbre relationnelle au calcul relationnel sûr

- Si R a pour attributs $A_1, \ldots, A_n, B_1, \ldots, B_m$, alors $\pi_{A_1,\ldots,A_n}(R)$ se traduit

$$9y_1, \ldots, y_m \; R(x_1, \ldots, x_n, y_1, \ldots, y_m)$$

  Important : ce sont les attributs qui ne sont *pas* projetés qui sont quantifiés.

  Exemple : si R a pour attributs A, B, $\pi_A(R)$ se traduit $9x_2 \; R(x_1, x_2)$

- $R \quad S$ se traduit $R(x_1, \ldots, x_n) \wedge S(y_1, \ldots, y_m)$ toutes les variables sont distinctes et le résultat aura donc $n + m$ attributs

Du l'algèbre relationnelle au calcul relationnel sûr

- Si R et S ont même ensemble d'attributs, alors $R \cap S$ se traduit $R(x_1, \ldots, x_n) \wedge S(x_1, \ldots, x_n)$
  (toutes les variables sont les mêmes et le résultat aura donc n attributs)

- Si R et S ont même ensemble d'attributs, alors $R - S$ se traduit $R(x_1, \ldots, x_n) \wedge \neg S(x_1, \ldots, x_n)$
  (toutes les variables sont les mêmes et le résultat aura donc n attributs)

Préliminaires à la traduction du calcul relationnel sûr dans l'algèbre relationnelle

- Domaine actif d'une relation : l'ensemble des constantes qui y occurrent.

  Exemple : le domaine actif de

R_1	A	B
	a_1	b_1
	a_2	b_2

  est fa_1, a_2, b_1, b_2g

- Calcul du domaine actif de R
  Soit R avec des attributs $A_1, \ldots, A_n$

$$ADOM(R) = \rho_{B \leftarrow A_1}(\pi_{A_1}(R)) \; [\; \ldots \; [\; \rho_{B \leftarrow A_n}(\pi_{A_1}(R))$$

- On construit une relation sur un unique attribut B
- De même on peut calculer

$$ADOM(R_1, \ldots, R_k) = ADOM(R_1) \; [\; \ldots \; [\; ADOM(R_k)$$

Du calcul relationnel sûr à l'algèbre relationnelle

- Une requête sûre sur les relations $R_1, \ldots, R_n$ ne peut produire d'élément hors de $ADOM(R_1, \ldots, R_n)$.
- I.e., pour une requête sûre Q,

$$ADOM(Q(R_1, \ldots, R_n)) \quad ADOM(R_1, \ldots, R_n)$$

- Raison : tous les éléments hors de $ADOM(R_1, \ldots, R_n)$ se valent, si l'un est dans la réponse alors tous le sont, et donc la requête n'est pas sûre.
- On traduit donc les requêtes du calcul relationnel évaluées dans $ADOM(R_1, \ldots, R_n)$ en requêtes de l'algèbre relationnelle
- Chaque formule du calcul relationnel $F(x_1, \ldots, x_n)$ est traduite en une expression E_F qui produit une relation avec n attributs.

Du calcul relationnel sûr à l'algèbre relationnelle : traduction

- Cas faciles (pour R avec attributs $A_1, \ldots, A_n$) :

$$R(x_1, \ldots, x_n) \quad \text{se traduit} \quad R$$

$$9x_1 R(x_1, \ldots, x_n) \quad \text{se traduit} \quad \pi_{A_2,\ldots,A_n} R$$

- Cas moins faciles :
  - Une condition $c(x_1, \ldots, x_n)$ se traduit

$$\sigma_c(ADOM \quad \ldots \quad ADOM)$$

    e.g., $x_1 = x_2$ se traduit $\sigma_{x_1 = x_2}(ADOM \quad ADOM)$
  - Une négation $: R(\bar{x})$ se traduit

$$(ADOM \quad \ldots \quad ADOM) \quad R$$

    i.e., on ne calcule que les tuples d'éléments *de la base de données* qui n'appartiennent pas à R

Du calcul relationnel sûr à l'algèbre relationnelle : traduction

- Le cas le plus di□cile : la disjonction
- Soit R et S avec deux attributs

$$Q(x,y,z) = R(x,y) \_ S(x,z)$$

- Son résultat a trois attributs et consiste en tous les tuples (x,y,z) tels que :
  - soit (x,y) *2* R et z *2* *ADOM*,
  - ou bien (x,z) *2* S et y *2* *ADOM*
- Le premier ensemble de tuples est simplement R *ADOM*
- Le second est plus complexe à définir :

$$\pi_{\#1,\#3,\#5}(\sigma_{\#1=\#4 \wedge \#2=\#5}(S \quad ADOM \quad S))$$

- Q est donc traduite comme suit :

$$R \quad ADOM \ [\ \pi_{\#1,\#3,\#5}(\sigma_{\#1=\#4 \wedge \#2=\#5}(S \quad ADOM \quad S))$$

($\#i$ = i-ème élément du produit cartésien $S \times ADOM \times S$, relation 5-aire.)

Les requêtes avec "pour tout" dans l'algèbre relationnelle

- Trouver les réalisateurs dont les films sont joués dans tous les cinémas.

$$\{real \mid \forall (ci; ti') \in Projection\ \exists ti; act(Projection(ci; ti) \wedge Film(ti; real; act))\}$$

- On définit :
  - $C_1 = \pi_{cinema}(P)$
  - $C_2 = \pi_{cinema,realisateur}(F \bowtie P)$

  (pour sauver de l'espace on va utiliser P pour projection et F pour Film)
- C_1 contient tous les cinémas, C_2 contient tous les réalisateurs avec les cinémas où leurs films passent.
- Notre requête est :

$$freal\ j\ 8ci\ 2\ C_1\ (ci, real)\ 2\ C_2)g$$

Requêtes avec "pour tout"

$$\{real \mid \forall ci \in C_1\ (ci, real) \in C_2\}$$

se réécrit

$$\{real \mid \neg(\exists ci \in C_1\ (ci, real) \notin C_2)\}$$

La réponse à la requête est donc

$$\pi_{realisateur}(F) \quad V$$

où $V = \{real \mid (\exists ci \in C_1\ (ci, real) \notin C_2)\}$

Les paires $(ci, real)$ qui ne sont pas dans C_2 sont

$$(C_1 \quad \pi_{realisateur}(F)) \quad C_2$$

D'où :

$$V = \pi_{realisateur}((C_1 \quad \pi_{realisateur}(F)) \quad C_2)$$

Requêtes avec "pour tout"

- Requête : trouver les réalisateurs dont les films passent dans tous les cinémas.
- On obtient donc :

$$\pi_{realisateur}(F) - \pi_{realisateur}((\pi_{cinema}(P) \times \pi_{realisateur}(F)) - \pi_{cinema,realisateur}(F \ ./ \ P))$$

- Beaucoup moins intuitif que la description logique de la requête.
- Les langages procéduraux sont loin d'être aussi compréhensibles que les langages déclaratifs...

Pour tout et la négation dans SQL

- Trouver les réalisateurs dont les films passent dans tous les cinémas.
- La façon SQL de dire ça : trouver tous les réalisateurs tels qu'il n'existe pas de cinéma où leurs films ne passent pas.

```
SELECT F1.realisateur
FROM Film F1
WHERE NOT EXISTS (SELECT P.cinema
                  FROM Projection P
                  WHERE NOT EXISTS (SELECT F2.realisateur
                                    FROM Film F2
                                    WHERE F2.titre=P.titre
                                        AND
                                    F1.realisateur=F2.realisateur))
```

Pour tout et la négation dans SQL

Même requête avec EXCEPT

```
SELECT F.realisateur
FROM Film F
WHERE NOT EXISTS (SELECT P.cinema
                  FROM Projection P
                    EXCEPT
                  SELECT P1.cinema
                  FROM Projection P1, Film F1
                  WHERE P1.titre=F1.titre
                  AND F1.realisateur=F.realisateur)
```

- Autres conditions : IN, NOT IN, EXISTS...

Pour tout et la négation dans SQL

- Deux mécanismes principaux : les sous requêtes et les expressions ensemblistes
- Sous-requêtes souvent plus naturelles
- Syntaxe de SQL pour $R \setminus S$:

  `R INTERSECT S`
- Syntaxe de SQL pour $R \quad S$:

  `R EXCEPT S`
- Trouver tous les acteurs qui

- ne sont pas réalisateurs :

```
SELECT acteur AS personne
FROM Film
 EXCEPT
SELECT realisateur AS personne
FROM Film;
```

- sont aussi réalisateurs :

```
SELECT acteur AS personne
FROM Film
 INTERSECT
SELECT realisateur AS personne
FROM Film;
```

Requêtes sans intersect et except

- Trouver tous les acteurs qui

- ne sont pas réalisateurs :

```
SELECT acteur  AS personne
FROM Film
 EXCEPT
SELECT realisateur AS personne
FROM Film;
```

- sont aussi réalisateurs :

```
SELECT acteur AS personne
FROM Film
 INTERSECT
SELECT realisateur AS personne
FROM Film;
```

- Requêtes alternatives (possiblement avec duplicats) :

- ne sont pas réalisateurs :

```
SELECT acteur
FROM Film
WHERE acteur NOT IN
 (SELECT realisateur
 FROM Film);
```

- sont aussi réalisateurs :

```
SELECT acteur
FROM Film
WHERE
acteur=realisateur;
```

Plus d'exemples de requêtes imbriquées : avec EXISTS et IN

Trouver les réalisateurs dont on joue les films au Champo.

```
SELECT F.realisateur
FROM Film F
WHERE EXISTS (SELECT *
              FROM Projection P
              WHERE F.titre=P.titre
              AND P.cinema='Le Champo');
```

```
SELECT F.realisateur
FROM Film F
WHERE F.titre IN (SELECT P.titre
              FROM Projection P
              WHERE P.cinema='Le Champo');
```

Plus d'exemples de requêtes imbriquées : avec NOT IN

Trouver les acteurs qui n'ont pas joué dans un film de Kubrick

```
SELECT F.acteur
FROM Film F
WHERE F.acteur NOT IN
          (SELECT F1.acteur
           FROM Film F1
           WHERE F1.realisateur='Kubrick');
```

La sous requête trouve les acteurs qui jouent dans des films de Kubrick, les trois lignes du haut prennent le complément de cet ensemble.

Trouver les acteurs qui n'ont joué que dans un seul film :

```
SELECT F1.acteur
FROM Film F1
WHERE F1.acteur NOT IN
                (SELECT F2.acteur
                FROM Film F2
                WHERE F.1=titre <> F2.titre);
```

Solution alternative :

```
SELECT F1.acteur
FROM Film F1
WHERE NOT EXISTS
      (SELECT *
      FROM Film F2
      WHERE F1.acteur=F2.acteur and
      F1.titre <> F2.titre);
```

Trouver les acteurs qui ont joué dans un film de Polanski mais pas dans un film de Kubrick :

```
SELECT F1.acteur
FROM Film F1
WHERE F1.acteur IN
      (SELECT F2.acteur
      FROM Film F2
      WHERE F2.realisateur='Polanski')
            AND
      F1.acteur NOT IN
      (SELECT F3.acteur
      FROM Film F2
      WHERE F3.realisateur='Kubrick');
```

Trouver les acteurs qui ont joué dans un film de Polanski mais pas dans un film de Kubrick :

```
SELECT F1.acteur
FROM Film F1
WHERE F1.acteur IN
      (SELECT F2.acteur
      FROM Film F2
      WHERE F2.realisateur='Polanski')
            AND
      F1.acteur NOT IN
      (SELECT F3.acteur
      FROM Film F2
      WHERE F3.realisateur='Kubrick');
```

Doublons

```
SELECT * FROM T1

A1   A2
---- ----
1    2
2    1
1    1
2    2

SELECT A1 FROM T1

A1
--
1
2
1
2
```

Doublons

- SELECT ne correspond pas exactement à l'opérateur de projection de l'algèbre relationnelle.
- La projection retourne l'ensemble $\{1, 2\}$
- SELECT conserve les doublons
- Comment omettre les doublons ? Utiliser SELECT DISTINCT

Gérer les doublons

- Jusqu'à présent dans l'algèbre relationnelle, on a opéré sur des ensembles. SQL, opère en fait sur des multi-ensembles, i.e., des ensemble pouvant contenir des doublons.
- Requière de petits ajustements
- La projection π ne retire plus les doublons :

$$\pi_A \left(\begin{array}{c|c} A & B \\ \hline a_1 & b_1 \\ a_2 & b_2 \\ a_1 & b_2 \end{array} \right) = \{a_1, a_2, a_1\}$$

  Ici a_1 apparaît deux fois.
- Il y a une opération spéciale d'élimination des doublons : elimination doublons($fa1, a2, a1g$) = $fa1, a2g$

Gérer les doublons : l'union

- L'opération d'union groupe deux multi-ensembles :

$$S = f1, 1, 2, 2, 3, 3g$$

$$T = f1, 2, 2, 2, 3g$$

$$S \ [\ T = f1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3g$$

  i.e., si *a* occure k fois dans S, et m fois dans T, alors *a* occure k + m fois dans $S \ [\ T$.
- Ceci ne correspond pas à l'opération `UNION` de SQL, qui élimine les doublons.
- Pour conserver les doublons, utiliser `UNION ALL` :

```
SELECT * FROM S
   UNION ALL
SELECT * FROM T;
```

Gérer les doublons : l'intersection

- L'opération d'intersection conserve le nombre d'occurrences minimal d'un élément :

$$S = f1, 1, 2, 2, 3, 3g$$

$$T = f1, 2, 2, 2, 3g$$

$$S \setminus T = f1, 2, 2, 3g$$

  i.e., si *a* occure k fois dans S, et m fois dans T, alors *a* occure min(k + m) fois dans $S \setminus T$.
- Ceci ne correspond pas à l'opération `INTERSECT` de SQL, qui élimine les doublons.
- Pour conserver les doublons, utiliser `INTERSECT ALL` :

```
SELECT * FROM S
   INTERSECT ALL
SELECT * FROM T;
```

Gérer les doublons : la différence

- L'opération de di□érence fonctionne comme suit :

$$S = f1, 1, 2, 2, 3, 3g$$

$$T = f1, 2, 2, 2, 3g$$

$$S \quad T = f1, 3g$$

  i.e., si *a* occure k fois dans S, et m fois dans T, alors *a* occure k - m fois dans $S \quad T$.

- Ceci ne correspond pas à l'opération `INTERSECT` de SQL, qui élimine les doublons.
- Pour conserver les doublons, utiliser `EXCEPT ALL` :

```
SELECT * FROM S
   EXCEPT ALL
SELECT * FROM T;
```

SQL n'est pas un langage de programmation

- Calculer 2+2 en SQL
- Etape 1 : il nous faut une table sur laquelle opérer :

```
CREATE TABLE Arbitraire (a int);
```

- 2+2 doit aller dans une clause SELECT. Il faut aussi lui donner un nom d'attribut.
- Essai :

```
SELECT 2+2 as X
FROM Arbitraire;
```

```
 X
-----------
  0 record(s) selected.
```

SQL n'est pas un langage de programmation

- Problème : il n'y avait pas de tuple dans `Arbitraire`
- Peuplons notre table :

```
INSERT INTO Arbitraire VALUES 1;
INSERT INTO Arbitraire VALUES 5;
SELECT 2+2 as X
FROM Arbitraire;
```

```
X
-----------
          4
          4

  2 record(s) selected.
```

SQL n'est pas un langage de programmation

- Il faut aussi éliminer les doublons...
- Et finalement :

```
SELECT DISTINCT 2+2 as X
FROM Arbitraire;
```

```
 X
-----------
           4

  1 record(s) selected.
```

Les pièges de l'ensemble vide

- Soit trois relations, S, T, R, sur le même attribut A.
- Requête : calculer $Q = R \setminus (S \cup T)$
- La requête suivante a l'air d'exprimer ça correctement :

```
SELECT R.A
FROM R, S, T
WHERE R.A=S.A OR R.A=T.A;
```

- Soit $R = S = \{1\}$, $T = \emptyset$. Alors $Q = \{1\}$, mais la requête SQL produit la table vide...
- Pourquoi ?

Les pièges de l'ensemble vide

- Soit trois relations, S, T, R, sur le même attribut A.
- Requête : calculer $Q = R \setminus (S \cup T)$
- La requête suivante a l'air d'exprimer ça correctement :

```
SELECT R.A
FROM R, S, T
WHERE R.A=S.A OR R.A=T.A;
```

- Soit $R = S = f1g$, $T = ;$. Alors $Q = f1g$, mais la requête SQL produit la table vide...
- Pourquoi ?
- Si T est vide, alors $R \quad S \quad T$ est vide aussi !

Plus sur la clause WHERE

- Une fois que nous avons des types (numériques, chaînes de caractères, etc), nous avons des opérations spécifiques aux types et donc des conditions de sélection spécifiques à ces types.

```
CREATE TABLE Finance (titre char(20), budget int,
recette int);
INSERT INTO Finance VALUES ('Shining', 19, 100);
INSERT INTO Finance VALUES ('Star wars', 11, 513);
INSERT INTO Finance VALUES ('Wild wild west', 170, 80);
```

Plus sur la clause WHERE

- Trouver les films qui ont perdu de l'argent

```
SELECT titre
FROM Finance
WHERE recette < budget;
```

- Trouver les films qui ont généré au moins 10 fois plus de recette que ce qu'ils ont coûté

```
SELECT titre
FROM Finance
WHERE recette > 10 * budget;
```

- Trouver le bénéfice généré par chaque film :

```
SELECT titre, recette - budget as profit
FROM Finance
WHERE recette - budget > 0;
```

Plus sur la clause WHERE

- Est-ce que Kubrick s'écrit avec "k" ou "ck" à la fin ?
- Pas besoin de se souvenir.

```
SELECT titre, realisateur
FROM Film
WHERE realisateur LIKE 'Kubr%';
```

- Est-ce que Polanski s'écrit avec "y" ou "i" à la fin ?
- Pas besoin de se souvenir.

```
SELECT titre, realisateur
FROM film
WHERE realisateur LIKE 'Polansk_';
```

Les comparaisons avec LIKE

- Motifs d'attributs avec LIKE
- Les motifs sont construit à partir de :
  lettres
  _, qui représente n'importe quelle lettre
  %, qui représente n'importe quelle sous-chaîne, dont l'ensemble vide
- Exemples :
  addresse LIKE '%Paris%'
  le motif '_a_b_' représente cacbc, aabba, etc
  le motif %a%b_' représente ccaccbc, aaaabcbcbbd, aba, etc

Les comparaisons avec LIKE

```
SELECT titre, realisateur
FROM film
WHERE realisateur LIKE 'Polansk_';
```

retourne l'ensemble vide.

- Parce que parfois x=y est vrai, alors que x LIKE y est faux !
- Raison : les espaces
- 'Polanski ' = 'Polanski ' est vrai, mais 'Polanski ' LIKE 'Polanski ' est faux.
- Si realisateur défini comme char(10), alors 'Polanski' est vraiment 'Polanski ' et ne correspond donc pas à 'Polanski_'.

Ajouter des attributs... vers les requêtes avec agrégation

```
ALTER TABLE Projection ADD COLUMN heure int DEFAULT 0;

UPDATE Projection
SET heure = 18
WHERE cinema='Le Champo' AND titre='Chinatown';

INSERT INTO Film VALUES ('Le Champo', 'Chinatown', 21);
```

ajoute l'attribut heure, et insère des valeurs pour cet attribut.

Plus d'une projection par film : utiliser d'abord `UPDATE`, puis `INSERT`.

Plus d'exemples avec de l'arithmétique

Requête : je veux voir un film de Lucas. Je ne peux pas y aller avant 19h, et je veux être sortie avant 23h.

Je veux voir : les cinémas et l'heure exacte à laquelle je sortirai, si mes conditions sont satisfaites.

```
SELECT P.cinema, P.heure + (F.duree/60.0) AS heurefin
FROM Projection P, Film F
WHERE F.titre=P.titre
   AND F.realisateur='Lucas'
   AND P.heure >= 19
   AND P.heure + (F.duree/60.0) < 23;
```

Requêtes d'agrégat simple

- Compter le nombre de tuples dans Film

```
SELECT COUNT(*)
FROM Film;
```

- Additionner la durée de tous les films

```
SELECT SUM(durée)
FROM Film;
```

Les doublons et l'agrégation

- Trouver le nombre de réalisateurs, approche naïve :

```
SELECT COUNT(realisateur)
FROM Film;
```

retourne le nombre de tuples dans Film.
Raison : SELECT ne supprime pas les doublons.

- Requête correcte :

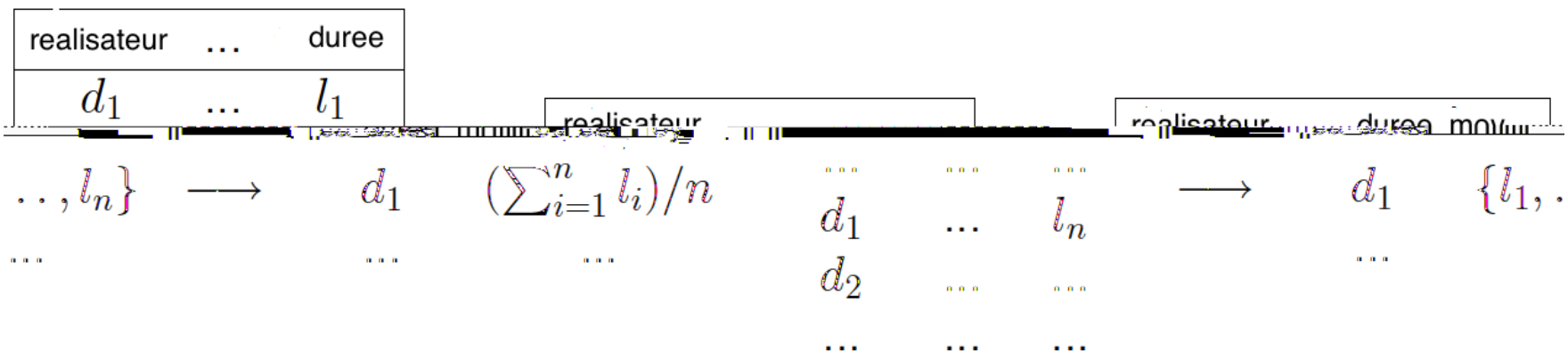
```
SELECT COUNT(DISTINCT realisateur)
FROM Film;
```

Agrégation et GROUP BY

Pour chaque réalisateur, retourner le temps moyen de ses films.

```
SELECT realisateur, AVG(duree) AS duree_moy
FROM Film
GROUP BY realisateur;
```

Comment GROUP BY fonctionne-t-il ?



(Attention : tous les attributs qui ne sont pas des agrégats et qui figurent dans la clause SELECT doivent figurer aussi dans la clause GROUP BY.)

Agrégation et doublons

Table	A1	A2	A3
	a	1	5
	a	1	2
	a	2	2
	a	2	3

```
SELECT A1, AVG(A3) as A4
FROM Table
GROUP BY A1;
```

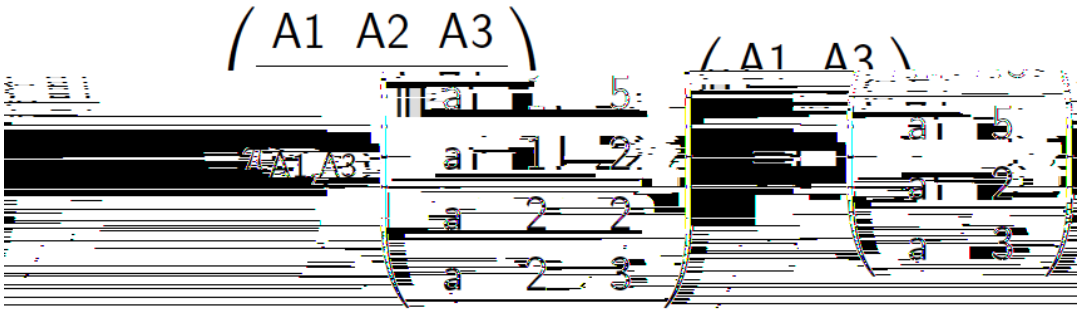
A1	A4
a	?

Aggrégation et doublons

Une approche : prendre toutes les valeurs de A3 et calculer leur moyenne

$$\frac{5+2+2+3}{4} = 3$$

Une autre approche : seuls les attributs A1 et A3 sont pertinents



$$\frac{5+2+3}{3} = \frac{10}{3}$$

Agrégation et doublons

- Approche SQL : toujours garder les doublons.
- La bonne réponse est donc 3.
- Attention, cependant :

  `SELECT AVG(A2) FROM Table;`

  retourne 1
- Raison : l'arrondi
- Solution : convertir comme nombre réel :

  `SELECT AVG(CAST (A2 AS REAL)) FROM Table;`

  retourne 1.5
- Syntaxe de `CAST`

  CAST (<attribut> AS <type>)

Plus sur les doublons

- Et si on veut éliminer les doublons avant de calculer l'agrégat ?
- Utiliser `DISTINCT`

  ```
  SELECT AVG(DISTINCT A3) FROM Table;
  ```

  donne 3, à cause de l'arrondi, mais

  ```
  SELECT AVG(DISTINCT CAST (A3 AS REAL)) FROM Table;
  ```

  produit 3.3333..., comme prévu
- Un truc pour convertir les entiers en réels :

  ```
  SELECT AVG(A3 + 0.0) FROM Table;
  ```

Autres fonctions d'agrégation

- `MIN` calcule la valeur minimum d'une colonne
- `MAX` calcule la valeur maximum d'une colonne
- `SUM` additionne tous les éléments d'une colonne
- `COUNT` compte le nombre de valeurs d'une colonne
- `MIN` et `MAX` produisent le même résultat peu importe les doublons
- `SUM` additionne tous les éléments d'une colonne donnée ; `SUM DISTINCT` additionne tous les éléments distincts d'une colonne donnée
- `COUNT` compte les éléments d'une colonne donnée ;
- `COUNT DISTINCT` compte les éléments distincts d'une colonne donnée

SUM, COUNT et doublons

- `SELECT COUNT(A3) FROM Table;` donne 4
- `SELECT COUNT(DISTINCT A3) FROM Table;` donne 3
- `SELECT SUM(A3) FROM Table;` donne 12
- `SELECT SUM(DISTINCT A3) FROM Table;` donne 10
- `SELECT MIN(A3) FROM Table;` et `SELECT MIN(DISTINCT A3) FROM Table;` donnent le même résultat.
- Idem pour `MAX`.

Sélections basées sur des résultats d'agrégation

- Trouver les réalisateurs et le temps moyen de leurs films, à condition qu'ils aient réalisé au moins un film de plus de 2 heures.
- Idée : calculer deux agrégats : AVG(duree) et MAX(duree) et choisir seulement les réalisateurs pour lesquels MAX(duree) > 120.
- Syntaxe de SQL pour ça : HAVING

```
SELECT realisateur, AVG(duree+0.0)
FROM Film
GROUP BY realisateur
HAVING MAX(duree) > 120;
```

Agrégation et jointures

- Les requêtes d'agrégation peuvent utiliser plus d'une relation.
- Pour tout cinéma montrant au moins un film de plus de 2 heures, trouver la durée moyenne des films qui y sont projetés.

```
SELECT F.cinema, AVG(CAST(F.duree AS REAL))
FROM Projection P, Film F
WHERE P.titre=F.titre
GROUP BY F.cinema
HAVING MAX(F.duree) > 120;
```

- Ce que ça veut dire : générer la jointure Film $\bowtie$ Projection, et sur cette jointure exécuter la requête d'agrégation qui calcule la moyenne.

Agrégation, jointures et doublons

- Les doublons peuvent donner lieu à des résultats inattendus.
- Deux tables :

R	A1	A2
	'a'	1
	'b'	2

S	A1	A3
	'a'	5
	'a'	7
	'b'	3

- Requête :

```
SELECT R.A1, SUM(R.A2)
FROM R, S
WHERE R.A1=S.A1 AND R.A1='a'
GROUP BY R.A1
HAVING MIN(S.A3) > 0;
```

- Résultat ?

Agrégation, jointures et doublons

- La table S n'est pas pertinente, et le résultat devrait être le même que celui de :

```
SELECT A1, SUM(A2)
FROM R
WHERE A1='a'
GROUP BY A1;
```

- Retourne ('a', 1)
- alors que la première requête retourne ('a', 2).

Agrégation, jointures et doublons

- Que se passe-t-il ?
- La requête construit d'abord la jointure $R \bowtie S$

R ⋈ S	A1	A2	A3
	'a'	1	5
	'a'	1	7
	'b'	2	3

- et exécute ensuite la partie agrégation sur la jointure :

```
SELECT A1, SUM(A2)
FROM R ⋈ S
WHERE A1='a'
GROUP BY A1
HAVING MIN(A3) > 0
```

- la réponse est donc ('a',2)

Agrégation, jointures et doublons

- Morale : attention aux doublons, même quand il n'y en a apparament pas.

- Pour retourner ('a',1), utiliser DISTINCT :

```
SELECT R.A1, SUM(DISTINCT R.A2)
FROM R, S
WHERE R.A1=S.A1 AND R.A1='a'
GROUP BY R.A1
HAVING MIN(S.A3) > 0;
```

Agrégats dans le WHERE

- Les résultats d'un agrégat peuvent être utilisés pour faire des comparaisons hors de la clause HAVING.
- Trouver les films plus longs que le film le plus long joué actuellement :

```
SELECT F.titre
FROM Film F
WHERE F.duree > (SELECT MAX(F1.duree)
                 FROM Film F1, Projection P
                 WHERE F1.titre=P.titre);
```

Agrégats dans le WHERE

- Attention à ne pas écrire :

```
SELECT F.titre
FROM Film F
WHERE F.duree > MAX(SELECT F1.duree
                    FROM Film F1, Projection P
                    WHERE F1.titre=P.titre);
```

qui est incorrect

- A la place on peut écrire :

```
SELECT F.titre
FROM Film F
WHERE F.duree > ALL(SELECT F1.duree
                    FROM Film F1, Projection P
                    WHERE F1.titre=P.titre);
```

Agrégats dans le WHERE

- De même :
- Trouver les films plus courts qu'au moins un film joué actuellement :

```
SELECT F.titre
FROM Film F
WHERE F.duree < (SELECT MAX(F1.duree)
                 FROM Film F1, Projection P
                 WHERE F1.titre=P.titre);
```

or

```
SELECT F.titre
FROM Film F
WHERE F.duree < ANY(SELECT F1.duree
                    FROM Film F1, Projection P
                    WHERE F1.titre=P.titre);
```

- Ici on utilise ANY et non ALL

ALL versus ANY

- <valeur> <condition> ALL (<requête>) est vrai si :
  - le résultat de <requête> est l'ensemble vide, ou
  - pour toute <valeur1> dans le résultat de <requête>, <valeur> <condition> <valeur1> est vrai.
- Par exemple,
  $5 > ALL(;)$ est vrai ;
  $5 > ALL(f1, 2, 3g$ est vrai ;
  $5 > ALL(f1, 2, 3, 4, 5, 6g$ est faux.

ALL versus ANY

- <valeur> <condition> ANY (<requête>) est vrai s'il existe une <valeur1> dans le résultat de <requête>, telle que <valeur> <condition> <valeur1> est vrai.
- Par exemple,
  $5 < ANY(;)$ est faux ;
  $5 < ANY(f1, 2, 3g$ est faux ;
  $5 < ANY(f1, 2, 3, 4, 5, 6g$ est vrai.

Agrégats dans le WHERE

- Toutes les comparaisons avec des résultats d'agrégat ne peuvent pas être remplacées par des comparaisons avec ANY et ALL.

- Est-ce qu'il y a un film dont la durée correspond au moins à 10% de la longueur totale de tous les autres films combinés ?

```
SELECT F.titre
FROM Film F
WHERE F.duree >= 0.1 * (SELECT SUM(F1.duree)
                        FROM Film F1
                        WHERE F1.titre <> F.titre);
```